Relabelling in Bayesian mixture models by pivotal units

Leonardo Egidi* Roberta Pappadà † Francesco Pauli†
Nicola Torelli†

Abstract

In this paper a simple procedure to deal with label switching when exploring complex posterior distributions by MCMC algorithms is proposed. Although it cannot be generalized to any situation, it may be handy in many applications because of its simplicity and very low computational burden. A possible area where it proves to be useful is when deriving a sample for the posterior distribution arising from finite mixture models, when no simple or rational ordering between the components is available.

1 Introduction

Label switching is a well-known and fundamental problem in Bayesian estimation of finite mixture models (McLachlan and Peel, 2000). The label switching problem arises when exploring complex posterior distributions by Markov Chain Monte Carlo (MCMC) algorithms because the likelihood of the model is invariant to the relabelling of mixture components.

Since there are as many maxima as there are permutations of the indices ($G!$), the likelihood has then multiple global maxima. This is a minor problem (if a problem at all) when we perform classical inference, since any maximum leads to a valid solution and inferential conclusions are the same regardless of which one is chosen.

On the contrary, invariance with respect to labels is a major problem when Bayesian inference is used. If the prior distribution is invariant with respect to the labelling as well as the likelihood, then the posterior distribution is multimodal.

To make inference on a parameter specific of a component of the mixture, a sample from the posterior that represent different modes would be inappropriate. An actual MCMC sample may or may not switch labels depending on the efficiency of the sampler. If the raw MCMC sampler randomly switches labels, then it is unsuitable for exploring the posterior distributions for component-related parameters.

*Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Italy, e-mail:egidi@stat.unipd.it
†Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'Bruno de Finetti', Università degli Studi di Trieste, Via Tigor 22, 34124 Trieste, Italy, e-mail: rpappada@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it

A range of solutions has been proposed to perform inference in presence of label switching (Frühwirth-Schnatter, 2001; Stephens, 2000), but for complex models most of the existing procedures are complex and computationally expensive.

A full solution entails obtaining valid samples for each parameter, and the methods in Section 5 are designed to relabel the raw Markov chains for this purpose. Simpler solutions are available if we do not need posterior samples for all the parameters.

In this paper a simple procedure based on the post MCMC relabelling of the chains to deal with label switching when exploring complex posterior distributions by MCMC algorithms is proposed.

As pointed out in Section 2.1, we can totally ignore the relabelling if the quantities of interest are label invariant. Besides the extreme case of label invariant quantities, we illustrate in Section 3 and 4 how to obtain a clustering and even a matrix of probabilities of units belonging to groups, using the raw MCMC sample without the need to fully relabel it. In Section 6 we propose a method which performs a relabelling starting from a suitable clustering of the samples, with the aim of using an MCMC sample to infer on the characteristics of the components in terms of both probabilities of each unit being in each group and the group parameters. The performance of the algorithm is explored via the simulation study discussed in Section 7 and a case study on real data is presented in Section 8. Section 9 concludes.

2 The relabelling problem

Prototypical models in which the labelling issue arises are mixture models, where, for a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ we assume

$$(Y_i|Z_i = g) \sim f(y; \mu_g, \phi),$$

where the Z_i, $i = 1, \ldots, n$, are i.i.d. random variables and

$$Z_i \in \{1, \ldots, G\}, \quad P(Z_i = g) = \pi_g.$$

The likelihood of the model is then

$$L(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\pi}, \phi) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f(y_i; \mu_g, \phi), \tag{1}$$

with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_G)$ component-specific parameters and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$ mixture weights. Equation (1) is invariant under a permutation of the indices of the groups, that is, if $(j_1, \ldots, j_G)$ is a permutation of $(1, \ldots, G)$ and $\boldsymbol{\pi'} = (\pi_{j_1}, \ldots, \pi_{j_G})$, $\boldsymbol{\mu'} = (\mu_{j_1}, \ldots, \mu_{j_G})$ are the corresponding permutations of $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$, then

$$L(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\pi}, \phi) = L(\boldsymbol{y}; \boldsymbol{\mu'}, \boldsymbol{\pi'}, \phi). \tag{2}$$

As a consequence, the model is unidentified with respect to an arbitrary permutation of the labels.

When Bayesian inference for the model is performed, if the prior distribution $p_0(\boldsymbol{\mu}, \boldsymbol{\pi}, \phi)$ is invariant under a permutation of the indices, that is $p_0(\boldsymbol{\mu}, \boldsymbol{\pi}, \phi) = p_0(\boldsymbol{\mu}', \boldsymbol{\pi}', \phi)$, then so is the posterior

$$p(\boldsymbol{\mu}, \boldsymbol{\pi}, \phi|\boldsymbol{y}) \propto p_0(\boldsymbol{\mu}, \boldsymbol{\pi}, \phi) L(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\pi}, \phi), \tag{3}$$

which is then multimodal with (at least) $G!$ modes. This implies that all simulated parameters should be switched to one among the $G!$ symmetric areas of the posterior distribution, by applying suitable permutations of the labels to each MCMC draw.

2.1 Relabelling and label switching in MCMC sampling

In the following we assume that we obtained an MCMC sample from the posterior distribution for model (1) with a prior which is labelling invariant. We denote as $\{[\theta]_h : h = 1, \ldots, H\}$ the sample for the parameter $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \phi)$. We assume that also the Z variable is MCMC sampled and denote as $\{[Z]_h : h = 1, \ldots, H\}$ the corresponding sample.

In principle, a perfectly mixing chain should visit the points $(\boldsymbol{\mu}, \boldsymbol{\pi}, \phi)$ and $(\boldsymbol{\mu}', \boldsymbol{\pi}', \phi)$ with the same frequency. A chain with these characteristics for a model with $G = 2$ and where $f(\cdot; \mu_g, \phi)$ is the Gaussian distribution with parameters μ_g and ϕ, $\mathcal{N}(\mu_g, \phi)$, is depicted in Figure 1(a), together with the posterior distribution for $\boldsymbol{\mu}$.

A chain with a less than perfect mixing may either concentrate on one mode of the posterior distribution (Figure 1(b)) or exhibit random switches (Figure 1(c)).

A naive, but effective, solution to the relabelling issue is to use a sampler which is inefficient with respect to the labelling – that is, it is unlikely to switch labels – but otherwise efficient (Puolamäki and Kaski, 2009). This can be an *ex post* solution, that is, we can ignore the relabelling issue if we verify that we obtained a chain where no switch occurred, but it is impractical in general terms since it is difficult to tune a sampler so that it is inefficient enough to avoid label switches but not too inefficient.

We note that the presence of label switches (or the whole issue of relabelling) is totally not relevant if the quantities we are interested in are invariant with respect to the labels, as is the case for a prediction of (y_1, y_2) (depicted in Figure 2, top row), or the inference for the parameter ϕ.

A particularly relevant example of invariant quantity is the probability of two units being in the same group, $c_{ij} = P(Z_i = Z_j|\mathcal{D})$, $i, j = 1, \ldots, n$, whose estimate based on the sample is

$$\hat{c}_{ij} = \frac{1}{H} \sum_{h=1}^{H} |[Z_i]_h = [Z_j]_h| \,. \tag{4}$$

The $n \times n$ matrix C with elements $\hat{c}_{ij}$ can be seen as an estimated similarity matrix between units, and the complement to one $\hat{s}_{ij} = 1 - \hat{c}_{ij}$ as a dissimilarity matrix (note that it is not a distance metric as $s_{ij} = 0$ does not imply that the units i and j are the same).

Relabelling becomes relevant when we are interested, directly or indirectly, in the features of the G groups, for example the posterior (and predictive) distributions of component-related quantities such as the difference $\mu_2 - \mu_1$ or the probability of each unit belonging to

each group, $q_{ig} = P(Z_i = g|\mathcal{D})$, whose MCMC estimate is

$$\hat{q}_{ig} = \frac{1}{H} \sum_{h=1}^{H} |[Z_i]_h = g| \,, \tag{5}$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, G$.

In Figure 2, bottom row, we depict the posterior distribution of $\mu_2 - \mu_1$ based on the samples $\{[\mu_2]_h - [\mu_1]_h : h = 1, \ldots, H\}$ obtained using the three chains. The first version is formally correct given that the model is not identified, but it is not able to tell us what is the average difference between the groups. The second version does answer to our question on the difference between the groups but is based on a very partial exploration of the posterior. The third version leads to an incorrect answer.

It is then clear that the raw MCMC sample can not be used to study the posterior distributions of component-related quantities such as μ_g or $P(Z_i = g|\mathcal{D})$. In order to study the posterior distributions of component-related quantities such as μ_g, we need to define a suitable method to permute the labels at each iteration of the Markov chain. Then, the new labels are such that different labels do refer to different components of the mixture.

3 Partitioning observations

A partition of the observations, meaning a point estimate of the group for each unit, can be easily obtained. Doing this, however, the issue of obtaining an estimate for groups features (posteriors of μ_g) or the probability of units belonging to each group ($\hat{q}_{ig}$ in Equation (5)) remains open. In fact, the usual difficulties related to clustering techniques apply (for instance, the groups depend on the choice of the distance). A partition can be also obtained by maximizing the posterior distribution, notwithstanding the fact that the maximum is not unique (there are $G!$ modes), since the maxima are equivalent any would be suitable.

Alternatively, the probabilities in Equation (4) can be used to derive a partition of observations by employing some clustering technique based on a suitable similarity matrix.

A more sophisticated option, see Fritsch and Ickstadt (2009), involves defining a distance between partitions, for example

$$d(\boldsymbol{z}^*, \boldsymbol{z}) = \sum_{i<k} d_1 \, |z_i^* \neq z_k| \, |z_i^* = z_k| + d_2 \, |z_i^* = z_k| \, |z_i^* \neq z_k| \,, \tag{6}$$

and then search for the partition which minimizes the expected distance with the true groups $\bar{\boldsymbol{z}}$, which means, if $d_1 = d_2 = 1$, find $\boldsymbol{z}^*$ which minimizes

$$E(d(\boldsymbol{z}^*, \bar{\boldsymbol{z}})|\mathcal{D}) = \sum_{i<k} \big| \, |z_i^* = z_k^*| - c_{ik} \big|, \tag{7}$$

where c_{ik} can be replaced by $\hat{c}_{ik}$.

Alternative distances between partitions may be used, for instance the Rand index $d_2(\boldsymbol{z}^*, \boldsymbol{z}) = 1 - d(\boldsymbol{z}^*, \boldsymbol{z}) \left(\binom{n}{2}\right)^{-1}$ or the adjusted Rand index (Hubert and Arabie, 1985).

Note that if the distance function is a linear operator then the following holds:

$$E(d(\boldsymbol{z}^*, \boldsymbol{z})|\mathcal{D}) = d\left(\boldsymbol{z}^*, E(\boldsymbol{z}|\mathcal{D})\right). \tag{8}$$

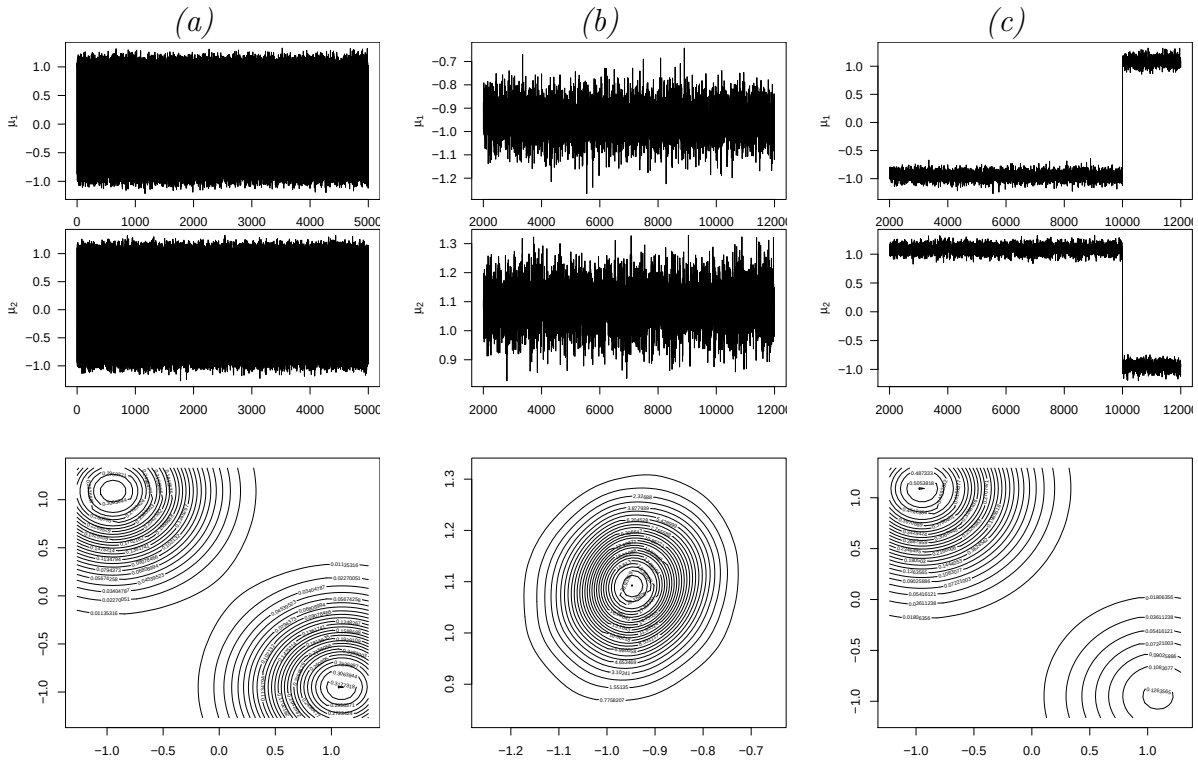


Figure 1: MCMC chains for $\boldsymbol{\mu}$ (top row) and estimated posterior for $\boldsymbol{\mu}$ where *(a)* a perfect mixing occurs (each of the two permutations is visited with equal frequency); *(b)* no switching is exhibited; *(c)* one random switch occurs.

The expectations in (7) or (8) can be obtained using the MCMC sample as

$$E(d(\boldsymbol{z}^*, \boldsymbol{z})|\mathcal{D}) = \frac{1}{H}\sum_{h=1}^{H} d(\boldsymbol{z}^*, [\boldsymbol{z}]_h). \tag{9}$$

The optimization should be done in the space of all possible partitions, since this can be very large, the authors suggest performing optimization on a suitable subset, reasonable alternatives being the set $\{[\boldsymbol{z}]_h\}$ or the set of clusterings resulting from different classical algorithms applied to the similarity matrix (4) (or the union of the two).

4 Obtaining probabilities of belonging to a group

Puolamäki and Kaski (2009) deal with the relabelling issue considering as an objective the $n \times G$ matrix with elements $q_{ig} = P(Z_i = g)$ ($\tilde{\beta}$ in their notation). This is obtained by maximizing the Bernoulli likelihood. The latter can be specified according to two alternative formulations. The first is one in Puolamäki and Kaski (2009), where the $HG \times n$ matrix Z' is such that

$$Z'_{ri} = 1 \text{ iff } [Z_i]_h = g \text{ where } r = G(h-1) + g,$$

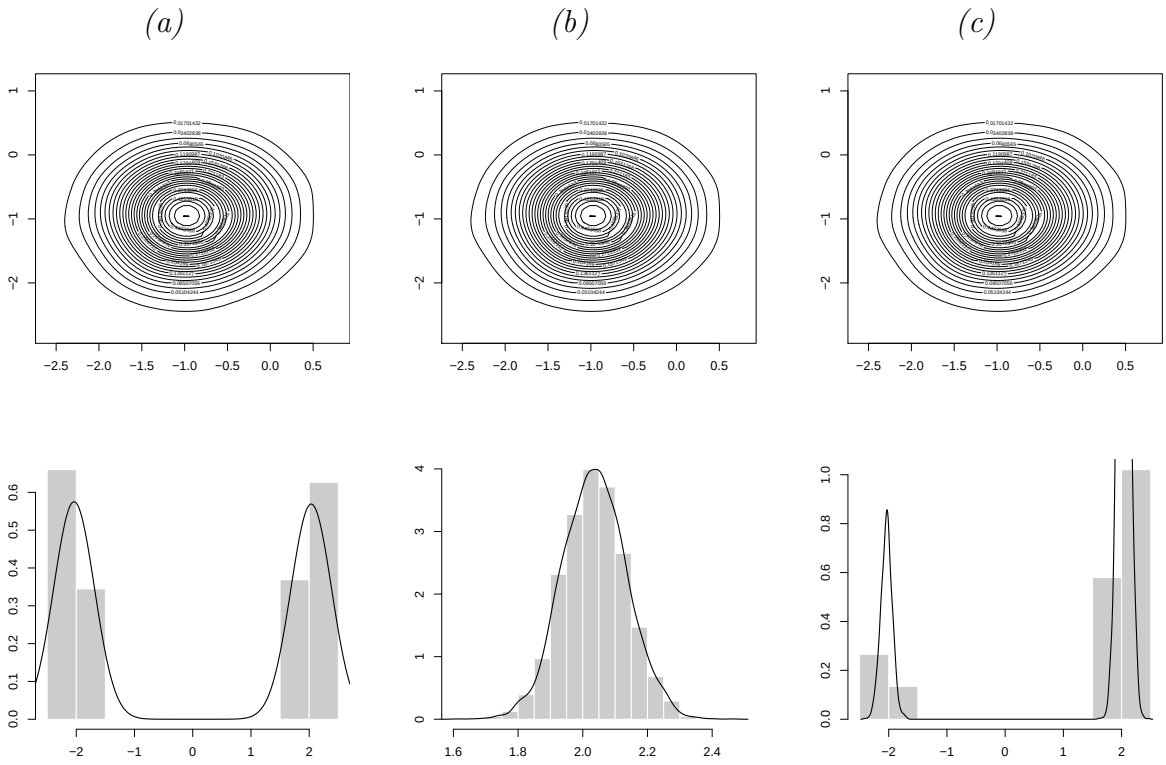


Figure 2: Estimated posteriors for (y_1, y_2) (top row) and $\mu_2 - \mu_1$ (bottom row) based on chain *(a)*, *(b)*, *(c)* of Figure 1.

and get

$$L = \prod_{r=1}^{R} \sum_{g=1}^{G} \prod_{i=1}^{n} q_{ig}^{Z'_{ri}} (1 - q_{ig})^{1 - Z'_{ri}}. \tag{10}$$

The above likelihood can also be written as

$$L = \sum_{h=1}^{H} \sum_{g=1}^{G} \prod_{i=1}^{n} q_{ig}^{|[Z_i]_h = g|} (1 - q_{ig})^{1 - |[Z_i]_h = g|}. \tag{11}$$

The intuitive idea behind this strategy is that if two units i_1 and i_2 often belong to the same group, that is, $Z'_{r,i_1} = Z'_{r,i_2}$ for many r, then they should be assigned to the same group, thus leading to a high value of $q_{i_1\bar{g}}$ and $q_{i_2\bar{g}}$ for some value of $\bar{g}$. Note that the likelihood above is itself labelling invariant, thus it has $G!$ maxima.

An EM algorithm is proposed to perform the optimization:

E step: for each row r (which represent a group in an iteration) and for each group obtain

$$\gamma_{rg} = \frac{\prod_{i=1}^{n} q_{ig}^{Z'_{ri}} (1 - q_{ig})^{1 - Z'_{ri}}}{\sum_{g=1}^{G} \prod_{i=1}^{n} q_{ig}^{Z'_{ri}} (1 - q_{ig})^{1 - Z'_{ri}}} = \frac{p((z_{1g}, \ldots, z_{ng})|\theta)}{\sum_{g=1} Gp((z_{1g}, \ldots, z_{ng})|\theta)};$$

M step: compute the mean of the Z'_{ri} with weights γ_{rg}

$$q_{ig} = \frac{\sum_{r=1}^{R} \gamma_{rg} Z'_{ri}}{\sum_{r=1}^{R} \gamma_{rg}} = \frac{\sum_{r:Z'_{ri}=1} \gamma_{rg}}{\sum_{r=1}^{R} \gamma_{rg}}.$$

Equivalently, the matrix Q can be found minimizing the cost function

$$\prod_{h=1}^{H} \sum_{\nu \in \mathcal{V}} \frac{1}{G!} q_{i\nu([Z_i]_h)}.$$

5 Relabeling methods

Relabelling means permuting the labels at each iteration of the Markov chain in such a way that the relabelled chain can be used to draw inference on component specific parameters. Loosely speaking we may say that the relabelled chain can be seen as a chain where no label switching has occurred or, in other words, the new labels are such that different labels do refer to different components of the mixture.

One method to perform the relabelling involves imposing identifiability constraints such as $\pi_1 < \pi_2 < \ldots < \pi_G$ or $\mu_1 < \mu_2 < \ldots < \mu_G$. Equivalently, this may be seen as a conditioning of the full (multimodal) posterior where the conditioning event is the identifiability constraint. Such a solution, although theoretically sound, may not be applicable when an obvious constraint does not exist and it may not work well if the components are not well separated (Stephens, 2000; Jasra et al., 2005).

It is worth noting that relabelling strategies may act during the MCMC sampling, and/or they may be used to post-process the chains. In general, those solutions which post-process

the chains are particularly convenient (since the issue can be ignored in performing the MCMC and then dealt with later). Generally, existing relabelling procedures select the permutation of the labels that minimizes a well defined distance between some components, such as pivots and classification probabilities, at each MCMC iteration. Papastamoulis (2016) provides the **label.switching** `R` package with a range of deterministic and probabilistic methods for performing relabelling: in Section 7 and 8 a comparison between some of these alternatives and our methodology will be provided.

5.1 Decision theoretic approach

A rather general decision theoretic framework for the relabelling problem is proposed by Stephens (2000). Such approach translates the problem to that of choosing an action a from a set of actions $\mathcal{A}$ where a loss function $\mathcal{L} : \mathcal{A} \times \Theta \to \mathbb{R}$ represents the loss we incur if we choose the action a and the true value of the parameter is θ.

The loss function makes sense if it is permutation invariant (remember that if we permute the parameter the model remains the same), we can obtain a permutation invariant loss function $\mathcal{L}$ from a non invariant one $\mathcal{L}_0$ by defining

$$\mathcal{L}(a;\theta) = \min_{\nu} \mathcal{L}_0(a;\nu(\theta)).$$

The action a is then chosen by minimizing the posterior expected loss

$$\mathcal{R}(a) = E(\mathcal{L}(a;\theta)|\mathcal{D}),$$

which can be approximated using the MCMC sample by

$$\begin{aligned}
\hat{\mathcal{R}}(a) &= \frac{1}{H}\sum_{h=1}^{H} \mathcal{L}(a;[\theta]_h) && (12)\\
&= \frac{1}{H}\sum_{h=1}^{H} \min_{\nu_h} \mathcal{L}_0(a;\nu_h([\theta]_h)) && (13)\\
&= \min_{\nu_1,\ldots,\nu_H} \left(\frac{1}{H}\sum_{h=1}^{H} \mathcal{L}_0(a;\nu_h([\theta]_h)) \right). && (14)
\end{aligned}$$

The action a can be the estimation of the parameter (or part of it) and the loss function may be a distribution to be fitted or an estimation error, the choice should be driven by the objective of inference. If the objective is the clustering of n units into G groups a reasonable action is reporting the $n \times G$ matrix $Q = [q_{ig}]$ where q_{ig} is the probability that the i-th unit belongs to the group g. A corresponding loss is then the distance, somehow measured (Stephens (2000) employs the Kullback-Leibler distance), between Q and its true value $P(\theta) = [p_{ig}(\theta)]$ where (for the toy example)

$$p_{ig}(\theta) = P(Z_i = g|\boldsymbol{y},\theta) = \frac{\pi_g f(y_i;\mu_g,\theta)}{\sum_j \pi_j f(y_i;\mu_j,\theta)}$$

The general algorithm for performing Stephens (2000) method is as follows

Start: from arbitrary permutations $\nu_1, \ldots, \nu_H$.

Step 1: obtain $a = \underset{a}{\text{argmin}} \sum_{h=1}^{H} \mathcal{L}_0(a; \nu_h([\theta]_h))$.

Step 2: obtain $\nu_h = \underset{\nu_h}{\text{argmin}} L_0(a; \nu_h([\theta]_h))$.

Note that step 2 entails n minimizations with respect to all the permutations ($G!$), Stephens (2000) points out the existence of efficient numerical algorithm if the loss function $\mathcal{L}_0$ can be written as $\mathcal{L}_0 = \sum_{g=1}^{G} \mathcal{L}_0^{(g)}(a; \pi_g, \mu_g, \phi)$.

A problem with this method might be the choice of the appropriate loss function/the dependence of the results on the loss function. Our method presented in Sect. 6 does not require a minimization step, and for this reason might be computationally appealing in many situations.

6 Pivotal method

Suppose that a partition of the observations in $\hat{G}$ groups, $\mathcal{G}_1, \ldots, \mathcal{G}_{\hat{G}}$ has been obtained as discussed in Section 3. As already pointed out, this may be enough for some purposes, but we may be interested in the probabilities $P(Z_i = g)$ and in the posteriors for groups parameters, μ_g.

Suppose that we can find $\hat{G}$ units, $i_1, \ldots, i_{\hat{G}}$, one for each group, which are (pairwise) separated with (posterior) probability one (that is, the posterior probability of any two of them being in the same group is zero). In terms of the matrix C, the $\hat{G} \times \hat{G}$ sub-matrix with only the row and columns corresponding to $i_1, \ldots, i_{\hat{G}}$ will be the identity matrix.

We then use the $\hat{G}$ units, called pivots in what follows, to identify the groups and to relabel the chains: for each $h = 1, \ldots, H$ and $g = 1, \ldots, \hat{G}$

$$[\mu_g]_h = [\mu_{[Z_{i_g}]_h}]_h; \tag{15}$$

$$[Z_i]_h = g \text{ for } i : [Z_i]_h = [Z_{i_g}]_h. \tag{16}$$

The availability of $\hat{G}$ perfectly separated units is crucial to the procedure, and it can not always be guaranteed. We now discuss three different circumstances under which the relabelling procedure is unsuitable

(i) the number of actual groups in the MCMC sample is higher than $\hat{G}$;

(ii) the number of actual groups in the MCMC sample is lower than $\hat{G}$;

(iii) the number of actual groups in the MCMC sample is equal to $\hat{G}$ but the pivots are not perfectly separated.

Let us first clarify what is meant by the number of actual groups. The model has G components, but some mixture components may be empty in the Markov chain, that is, it may happen that $\#\{g : [Z_i]_h = g \text{ for some } i\} < G \, \forall h$. By actual number of groups we mean the number of non empty groups, G_0 in what follows. It is then clear that the Markov chain does not have informations on more than G_0 groups.

We also note that the number of non empty groups may vary with iterations, let

$$[G]_h = \#\{g : [Z_i]_h = g \text{ for some } i\}.$$

Consider now the set $\mathcal{H}_1 \subset \{1, \ldots, H\}$ of iterations where $[G]_h > \hat{G}$; some units and groups will then have no available pivot. These units will not be attributed any group by performing (16). Thus for these units

$$\sum_{g=1}^{\hat{G}} \hat{P}(Z_i = g) = \sum_{g=1}^{\hat{G}} \hat{q}_{ig} = \sum_{g=1}^{\hat{G}} \frac{1}{H} \sum_{i=1}^{H} |[Z_i]_h = g| < 1.$$

We suggest cancelling those iterations of the chains where this occur, that is, the final –partial– chain is a sample from the posterior conditional on having at most $\hat{G}$ non empty groups.

Consider now the set $\mathcal{H}_2 \subset \{1, \ldots, H\}$ of iterations where $[G]_h < \hat{G}$; if $h \in \mathcal{H}_2$, $[Z_{i_k}]_h = [Z_{i_s}]_h$ for some pivots i_k, i_s. As a consequence, $\hat{c}_{hk} > 0$: the pivots are not perfectly separated. The procedure in (15) and (16) can not be performed (it is not well defined), so also in this case we will have to cancel the corresponding part of the chain. Finally, consider the set

$$\mathcal{H}_3 = \{h : \exists k, s \text{ s.t. } [Z_{i_k}]_h = [Z_{i_s}]_h\}$$

that is, the set of iterations where (at least) two pivots are put in the same group. Note that $\mathcal{H}_2 \subset \mathcal{H}_3$ but $\mathcal{H}_3$ may be larger. The same provision as above applies, we need to get rid of this part of the chain. In the end, we will relabel the chain with iterations

$$\mathcal{H}_0 = \{1, \ldots, H\} \backslash (\mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3) \tag{17}$$

which can be considered a sample from the posterior distribution conditional on (i) there being exactly $\hat{G}$ non empty groups, (ii) the pivots falling into different groups.

6.1 Pivots identification

A relevant issue is how to identify the pivots, noting that perfectly separated units may not exist and that, even if they exist, we may not be able to find them since the set of all possible choices is too big to be fully searched.

The general method we put forward is to select a unit for each group according to some criterion, for instance for group g containing units $\mathcal{G}_g$ we may chose $\bar{i} \in \mathcal{G}_g$ that maximizes one of the quantities

$$\max_{j \in \mathcal{G}_g} c_{\bar{i}j}, \quad \sum_{j \in \mathcal{G}_g} c_{\bar{i}j}, \quad \sum_{j \in \mathcal{G}_g} c_{\bar{i}j} - \sum_{j \notin \mathcal{G}_g} c_{\bar{i}j}; \tag{18}$$

or minimizes one of the quantities

$$\min_{j \in \mathcal{G}_g} c_{\bar{i}j}, \quad {}_{j \notin \mathcal{G}_g} c_{\bar{i}j}, \quad \sum_{j \notin \mathcal{G}_g} c_{\bar{i}j}. \tag{19}$$

We introduce a further method, which we call *Maxima Units Search* (hereafter MUS), that turns out to be suitable in case of a low number of mixture component, e.g., $G = 3, 4$.

This procedure differs from the others in the strategy for detecting pivots, since it does not rely upon a maximization/minimization step but it identifies those units satisfying a proper search within the estimated similarity matrix C (see the Appendix for more details on the MUS procedure).

The quality of the choice of pivotal units by the proposed methods is measured by the probability of the conditioning event

$$\frac{1}{H}\#\mathcal{H}_0. \tag{20}$$

estimated by the (original, raw) MCMC sample. It is worth noting that the idea of solving the relabelling issue by fixing the group for some units dates back to Chung et al. (2004), who, however, gave no indication on how to choose the units. Also, since they suggest imposing such a restriction in the MCMC, there is no measure of the extent to which it influences the result (of the extent to which it is informative if we interpret it as a prior information). We note, however, that Chung et al. (2004) may be very interesting when a set of units which are to be attributed to different groups can be defined exogenously.

Another related idea is put forward by Yao and Li (2014), who propose finding a reference labelling, that is, a clustering for the sample (for example, the posterior mode), and then relabel each iteration by minimizing some distance from the reference labelling. The general idea is similar to the one we suggest, but it is more computationally demanding because of the required minimizations, on the other hand it avoids the need to condition on the pivots being separated. We can argue, however, that the latter is not a big drawback of our proposal since its effects can be measured and is likely to be small in many practical instances.

7 Simulation study

The aim of this section is to evaluate the performance of the pivotal method introduced before. In particular, our goal is to investigate the behaviour of the proposed solution for dealing with label switching in different simulated scenarios. For this purpose, we focus on data simulated from a mixture of non-equally weighted mixtures of bivariate Gaussian distributions with unequal covariance matrices, so that the generated components may result in overlapping clusters. Specifically, the simulation scheme consists in the following steps.

(i) Simulate n values $Y_1, \ldots, Y_n$, from a mixture of mixtures of bivariate Gaussian distributions, where

$$(Y_i|Z_i = g) \sim \sum_{s=1}^{2} p_{gs}\,\mathcal{N}_2(\mu_{gs}, \Sigma_s). \tag{21}$$

That is, conditionally on being in group $g \in \{1, \ldots, G\}$, $\mathbf{y}_i$ is picked out from one of two possible Gaussian distributions with weights, means and covariances p_{gs}, μ_{gs} and Σ_s, $s = 1, 2$, respectively. The likelihood of the model is then

$$L(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\pi}, \Sigma) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \left(\sum_{s=1}^{2} p_{gs}\,\mathcal{N}_2(\mu_{gs}, \Sigma_s) \right).$$

	Scenario A	Scenario B	Scenario C
μ_{1s}	(25,0)	(-10,-10)	(-10,-10)
μ_{2s}	(60,0)	(20,-10)	(20,-10)
μ_{3s}	(0,20)	(-10,20)	(5,5)
μ_{4s}	(50,20)	(20,20)	(5,25)

Table 1: Two-dimensional mean vectors used as input for the three illustrated scenarios A, B and C, with number of mixture components $G = 4$.

(ii) Obtain an MCMC sample which effectively explores all modes of the posterior distribution.

(iii) Estimate the $n \times n$ similarity matrix C with elements $c_{ij} = P(Z_i = Z_j|\mathcal{D})$, $i, j = 1, \ldots, n$, by Equation (4).

(iv) Apply a suitable clustering technique based on the estimated dissimilarity matrix with elements $\hat{s}_{ij} = 1 - \hat{c}_{ij}$ and obtain a partition of the observations in $\hat{G}$ groups with units $\mathcal{G}_g, g = 1, \ldots, \hat{G}$.

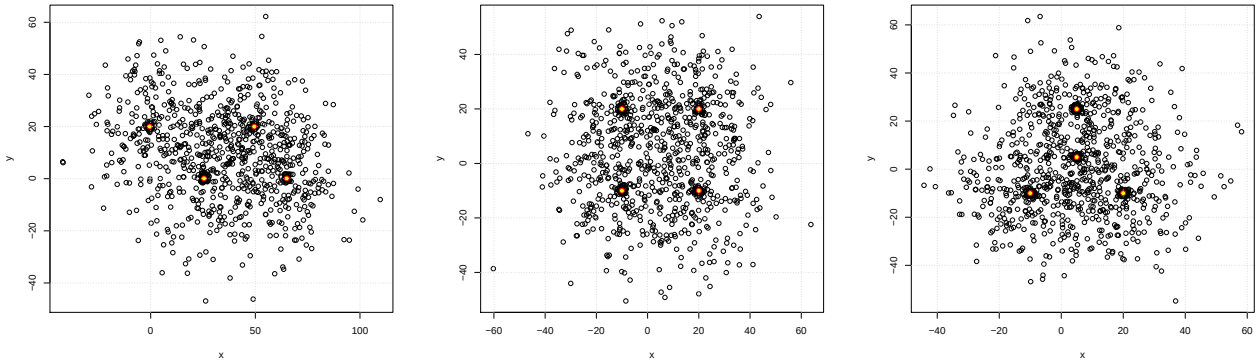
(v) Detect the pivots, one for each group, according to one criterion among the ones discussed before.

(vi) If necessary, discard those iterations of the chains belonging to $\mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3$ (see Section 6) and relabel the resulting chain with iterations in $\mathcal{H}_0$ (see Equation (17)) via (15) and (16).

In the following, a sample size of $n = 1000$ and $G = 4$ components are considered. For $g = 1, \ldots, 4$, we set $\pi_g = 1/4$, $p_{g1} = 0.2$, $p_{g2} = 0.8$, and $\Sigma_1 = \mathbf{I}_2$, $\Sigma_2 = 200\,\mathbf{I}_2$, being $\mathbf{I}_2$ the 2×2 identity matrix. We generate our simulated data from model (21) (see Figure 3) using the input means reported in Table 1 and obtain an MCMC sample by considering $H = 3000$ iterations.

We proceed following points (i)-(vi) described above. As a remark, two different clustering strategies are applied on the dissimilarities $\hat{s}_{ij}$ in order to obtain $\hat{G}$ clusters of observations, namely the agglomerative and partitioning hierarchical clustering. Both methods only require a distance or a dissimilarity matrix as input and return a set of nested clusters that are organized as a tree. The former starts with the points as individual clusters and, at each step, merges the closest pair of clusters, according to some criterion to compute cluster proximity; the latter starts with one, all-inclusive cluster and, at each step, splits a cluster until only singleton clusters of individual points remain.
We observe that the two algorithms provide very similar results in terms of the resulting clusters, and do not affect the performance of the relabelling procedure. Therefore, for the sake of illustration, we restrict to agglomerative hierarchical clustering, where the so-called *complete linkage* is adopted as a common criterion for the computation of the dissimilarity between two clusters, since it is less susceptible to noise and outliers than other linkages.

Figures 4–6 display the results of the agglomerative hierarchical clustering on simulated data from scenarios A, B and C, respectively. In each chart of Figures 4–6 a different method



(a) Scenario A (b) Scenario B (c) Scenario C

Figure 3: Illustration of a simulated sample of size $n = 1000$ from model (21) with $G = 4$ components, according to three different scenarios. The input means coordinates are reported in Table 1.

for identifying the pivots is adopted ((a)-(g)), and the selected units are marked with red points on the plots. Recall that, by definition, the pivots are perfectly (pairwise) separated units. Therefore, the performance of the seven different identification methods will be higher as the posterior probability of any two of them being in the same group is closer to zero. As can be noticed, panels (b), (e), (f) and (g) seem to provide an accurate choice of the pivots in all situations, since they are clearly well separated and suitable as representative units for each group. This is not a negligible issue in terms of the relabelling performance, which is strongly affected by the choice of such $\hat{G}$, by virtue of Equation (15). The estimated proportions of relabelled iterations based on 100 simulated samples are reported in Table 2. As expected, better performances in terms of pivot selection are likely to reflect into a higher proportion of relabelled iterations in almost all situations.

Coherently with the considerations drawn from Figures 4–6, methods (b), (e) and (f) register the highest chain proportions (less than 1% of the iterations is discarded) for both scenarios A and B. Method (c) seems to have the worst performance regardless of the considered scenario, in particular, for scenario C the chain keeps only about 8% of the original iterations. The fact that the third simulated scenario shows globally less satisfactory results is not surprising. In fact, the input means are so close each other that the cluster algorithms may fail in recognizing the true data partition, thus reflecting on the quality of the choice of the pivotal units. However, (e) and (f) criteria and MUS algorithm are preferable to the others.

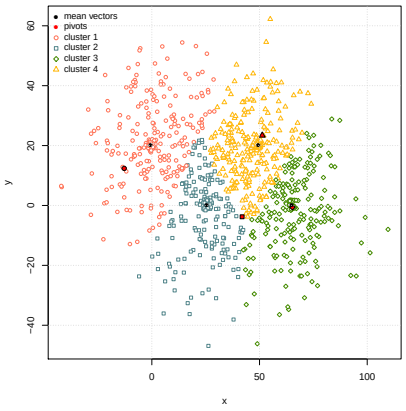
In order to compare the proposed methodology with other relabelling algorithms, in the task of estimating the means of the mixture components, we consider the Puolamäki and Kaski procedure (Puolamäki and Kaski, 2009) and three other methods implemented in the **label.switching** package (Papastamoulis, 2016). In Figure 7 the median estimates of relabelled group means are plotted for a simulated example from scenario B and four alternative methods. As can be seen, our relabelling procedure seems to provide very accurate estimates of group means. Similar results are achieved by ECR-iterative-1, ECR and DATA-BASED,

while Puolamäki and Kaski algorithm appears not to yield reliable estimates for the group means.

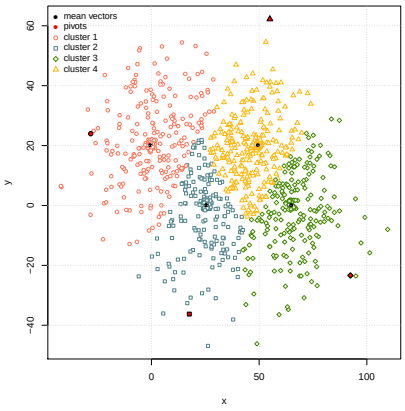
In Table 3 are reported the mean square errors of the relabelled estimates, obtained as mean over $B = 100$ macro-replications of the Euclidean distances between the input means and the corresponding estimates, according to scenarios A, B and C. In all scenarios the highest mean square errors are obtained for method (c), for each component of the mixture. Criteria (b), (e) and (f) give very similar results, and the MUS algorithm outperforms all other pivotal methods for three cases in Scenario A and two in Scenario B. ECR-iterative-1 performances in terms of mean square errors are comparable with our algorithm in most cases, although it shows to give lower errors in the third scenario, while Puolamäki and Kaski algorithm seems to provide poor estimates in all situations (see also Figure 7).

	(a)	(b)	(c)	(d)	(e)	(f)	(g)
Scenario A	0.475	0.993	0.124	0.506	0.993	0.993	0.313
Scenario B	0.519	0.998	0.101	0.707	0.998	0.998	0.995
Scenario C	0.139	0.300	0.079	0.267	0.368	0.507	0.374

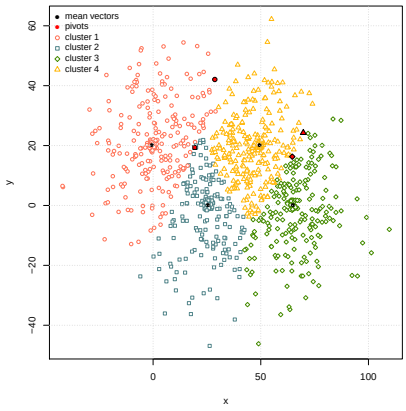
Table 2: Estimated proportion of relabelled iterations (see Equation (20)), over 100 macro-replications, based on the original MCMC sample, according to Scenario A, B and C. The observations are clustered according to agglomerative hierarchical algorithm.



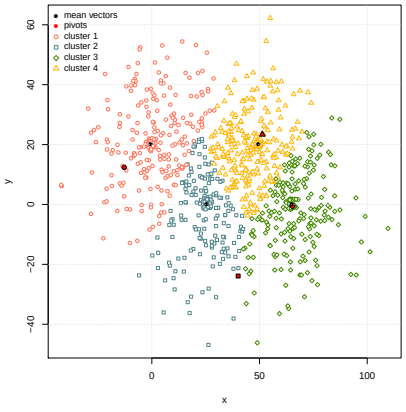
(a) $\max_{\bar{i}} \left(\max_{j \in \mathcal{G}_g} c_{\bar{i}j}\right)$



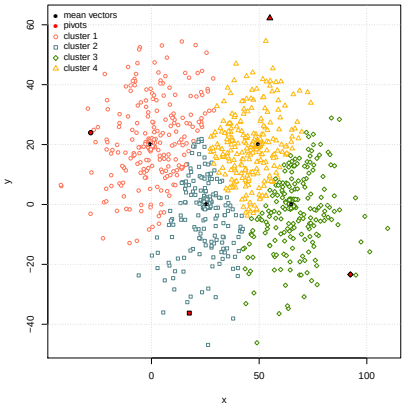
(b) $\max_{\bar{i}} \sum_{j \in \mathcal{G}_g} c_{\bar{i}j}$



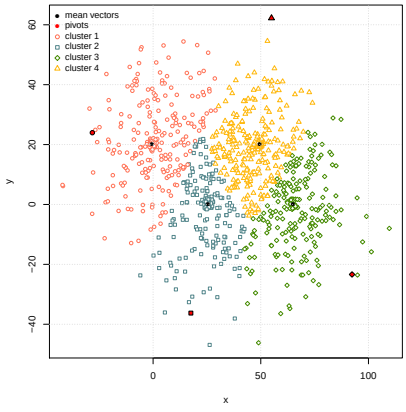
(c) $\min_{\bar{i}} (\min_{j \in \mathcal{G}_g} c_{\bar{i}j})$



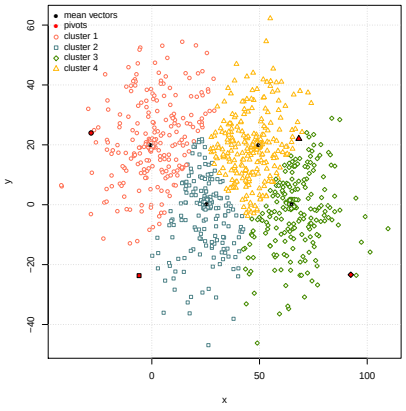
(d) $\min_{\bar{i}} (\min_{j \notin \mathcal{G}_g} c_{\bar{i}j})$



(e) $\min_{\bar{i}} \sum_{j \notin \mathcal{G}_g} c_{\bar{i}j}$

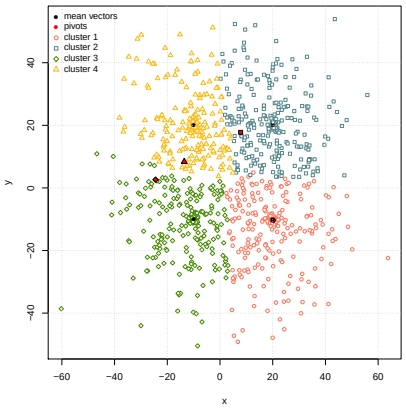


(f) $\max_{\bar{i}} \left(\sum_{j \in \mathcal{G}_g} c_{\bar{i}j} - \sum_{j \notin \mathcal{G}_g} c_{\bar{i}j}\right)$

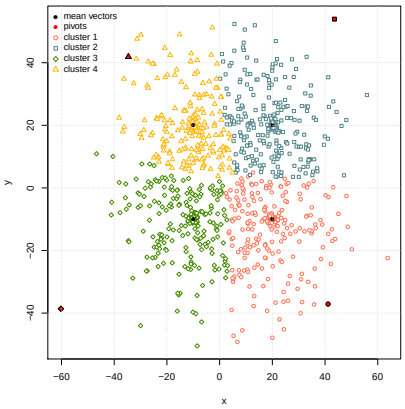


(g) MUS algorithm

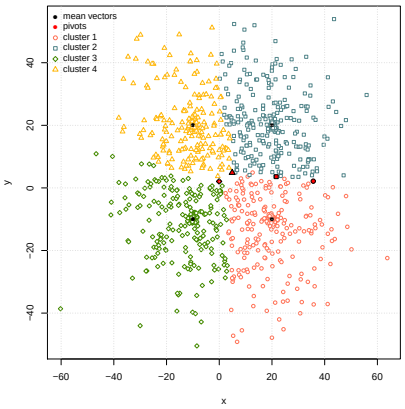
Figure 4: Simulated sample of size $n = 1000$ from Scenario A (see Table 1) clustered according to agglomerative hierarchical algorithm. The pivotal units are identified by adopting methods (a)–(g).



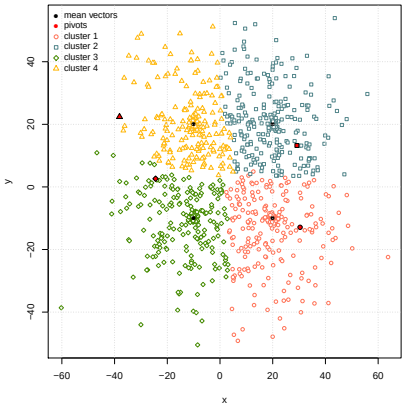
(a) $\max_{\bar{i}} \left(\max_{j \in \mathcal{G}_g} c_{\bar{i}j} \right)$



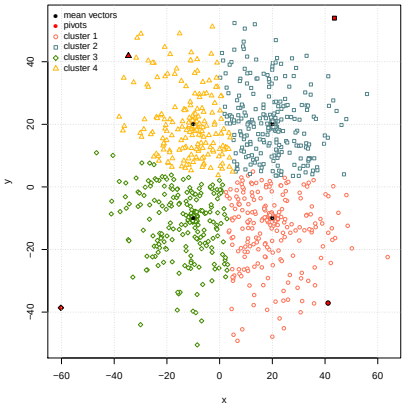
(b) $\max_{\bar{i}} \sum_{j \in \mathcal{G}_g} c_{\bar{i}j}$



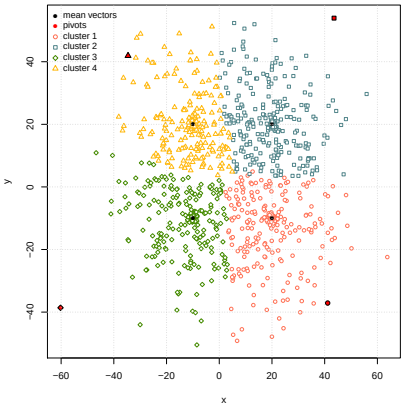
(c) $\min_{\bar{i}} (\min_{j \in \mathcal{G}_g} c_{\bar{i}j})$



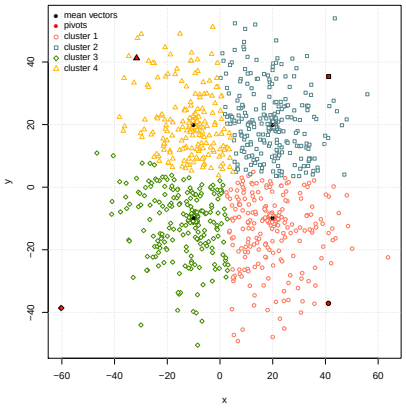
(d) $\min_{\bar{i}} (\min_{j \notin \mathcal{G}_g} c_{\bar{i}j})$



(e) $\min_{\bar{i}} \sum_{j \notin \mathcal{G}_g} c_{\bar{i}j}$

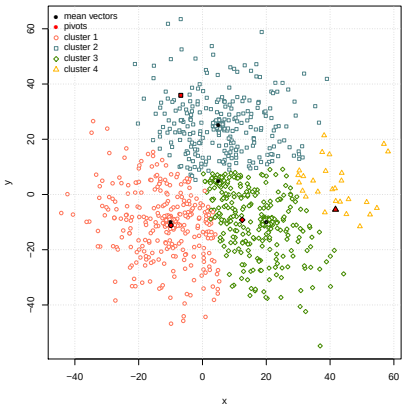


(f) $\max_{\bar{i}} \left(\sum_{j \in \mathcal{G}_g} c_{\bar{i}j} - \sum_{j \notin \mathcal{G}_g} c_{\bar{i}j} \right)$

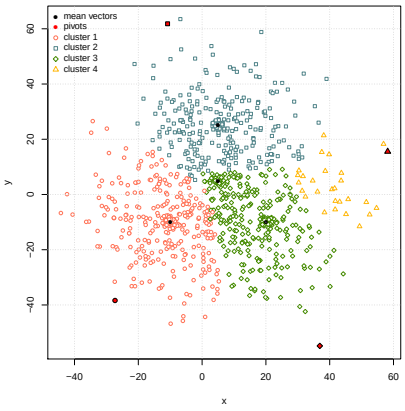


(g) MUS algorithm

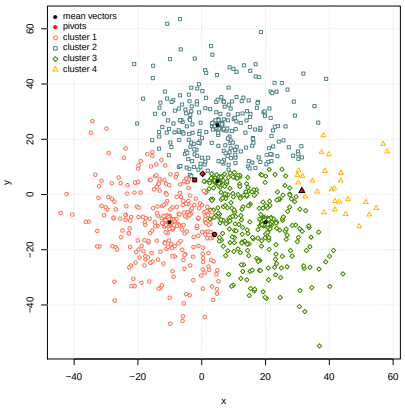
Figure 5: Simulated sample of size $n = 1000$ from Scenario B (see Table 1) clustered according to agglomerative hierarchical algorithm. The pivotal units are identified by adopting methods (a)–(g).



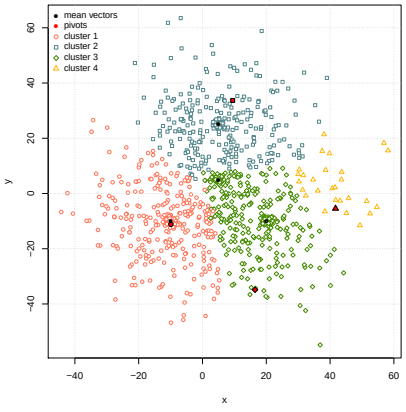
(a) $\max_{\bar{i}} \left(\max_{j \in \mathcal{G}_g} c_{\bar{i}j}\right)$



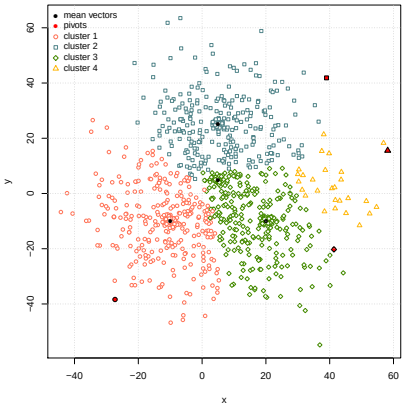
(b) $\max_{\bar{i}} \sum_{j \in \mathcal{G}_g} c_{\bar{i}j}$



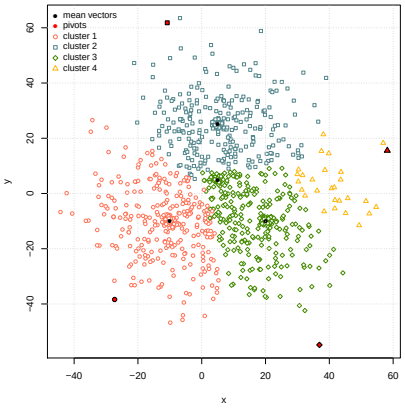
(c) $\min_{\bar{i}} (\min_{j \in \mathcal{G}_g} c_{\bar{i}j})$



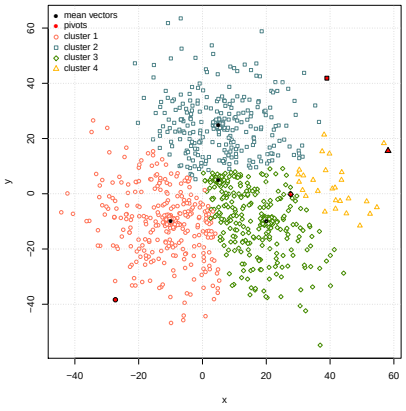
(d) $\min_{\bar{i}} (\min_{j \notin \mathcal{G}_g} c_{\bar{i}j})$



(e) $\min_{\bar{i}} \sum_{j \notin \mathcal{G}_g} c_{\bar{i}j}$



(f) $\max_{\bar{i}} \left(\sum_{j \in \mathcal{G}_g} c_{\bar{i}j} - \sum_{j \notin \mathcal{G}_g} c_{\bar{i}j}\right)$



(g) MUS algorithm

Figure 6: Simulated sample of size $n = 1000$ from Scenario C (see Table 1) clustered according to agglomerative hierarchical algorithm. The pivotal units are identified by adopting methods (a)–(g).

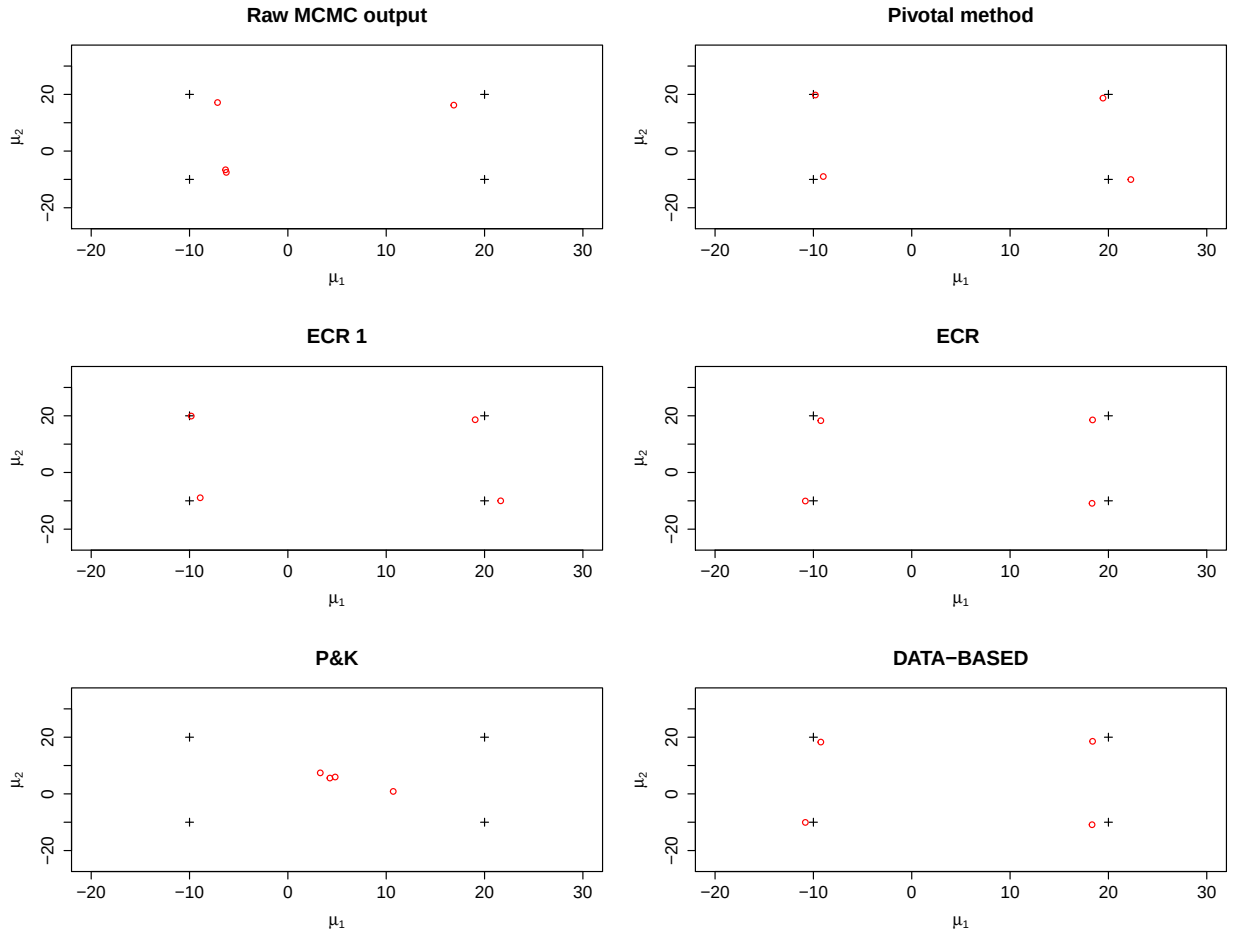


Figure 7: Scenario B. Crosses are input means, red points are the median values of relabelled estimates. *(Top left)* Raw MCMC sample for $\mu_g, g = 1, ..., 4$. *(Top right)* Reordered MCMC sample of Pivotal Method resulting from agglomerative hierarchical clustering and MUS algorithm. Reordered MCMC sample according to methods ECR-iterative-1, ECR, Puolamäki and Kaski and DATA-BASED.

Scenario A	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$	$\boldsymbol{\mu}_4$
(a)	13.8078	1.6724	2.0158	10.8232
(b)	13.7064	1.6104	1.9814	9.1846
(c)	22.8786	7.5307	6.7326	14.5896
(d)	14.0215	1.6619	1.9951	11.2910
(e)	13.7301	1.6264	1.8889	9.2900
(f)	13.7794	1.6723	1.8979	9.2897
MUS	**12.5787**	**1.5531**	**1.7919**	9.6220
ECR-1	13.6403	1.6605	1.9015	**8.8085**
P & K	25.5940	15.5229	15.1522	27.2411
Scenario B	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$	$\boldsymbol{\mu}_4$
(a)	1.4066	1.6251	1.5984	1.5884
(b)	1.4123	1.6005	1.5737	1.5419
(c)	4.8496	4.3588	4.8097	5.2142
(d)	1.4096	1.5961	1.5729	1.5403
(e)	1.4127	1.6003	1.5736	1.5417
(f)	1.4121	1.5982	1.6192	1.5420
MUS	**1.4070**	**1.5877**	1.5728	1.5437
ECR-1	1.4129	1.5984	**1.5717**	**1.5429**
P & K	18.4657	18.6185	18.6796	19.0404
Scenario C	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$	$\boldsymbol{\mu}_4$
(a)	9.1013	9.9974	8.4766	19.3288
(b)	6.9196	7.8994	8.7700	14.1766
(c)	11.6894	10.8810	8.8252	22.6435
(d)	7.7730	9.1701	9.1987	16.4153
(e)	7.6160	7.1054	10.2073	13.2589
(f)	7.1992	7.1643	9.4728	15.2713
MUS	6.7458	7.5579	9.7924	14.8356
ECR-1	**6.4891**	**6.7234**	8.4472	**9.3649**
P & K	17.5726	16.8717	**3.4988**	20.2620

Table 3: Mean squared error $\frac{1}{B}\sum_{j=1}^{B}||\boldsymbol{\mu}_{gs}^{(j)} - \hat{\boldsymbol{\mu}}_{gs}^{(j)}||$ computed for $B = 100$ macro-replications, of the estimates of the mean vector components $\boldsymbol{\mu}_{gs}$, $g = 1, ..., 4, s = 1, 2$, according to criteria (a)–(f) and MUS of pivotal method, Puolamäki and Kaski (P&K) and ECR-1 algorithms.

8 Case study

Fishery dataset, originally taken from Titterington et al. (1985) and used by Papastamoulis (2016) for comparing different relabelling procedures, consists of $n = 256$ snapper length measurements. In Figure 8 the histogram of the lengths is shown. In this section, we apply our method to these data and test its efficiency, comparing the results with some methods from the **label.switching** package. We use a Gaussian mixture with $G = 5$ components as suggested by Papastamoulis (2016), that is:

$$y_i \sim \sum_{g=1}^{G} p_g \mathcal{N}(\mu_g, \sigma_g^2), \quad i = 1, ..., n \tag{22}$$

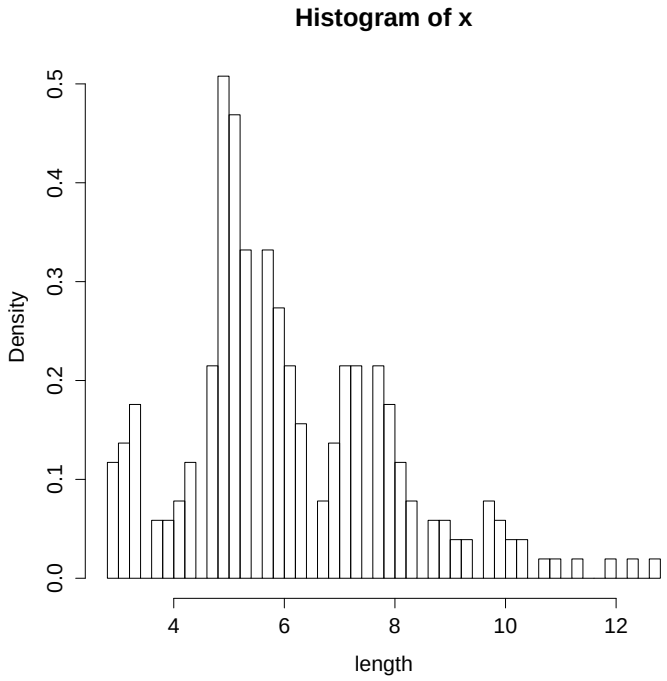


Figure 8: Histogram of fishery data. On x-axis the snapper length measurements

We set up a Gibbs sampling through the **bayesmix** `R` package (Grün, 2011), with $H = 11000$ iterations and a burn-in period of 1000.

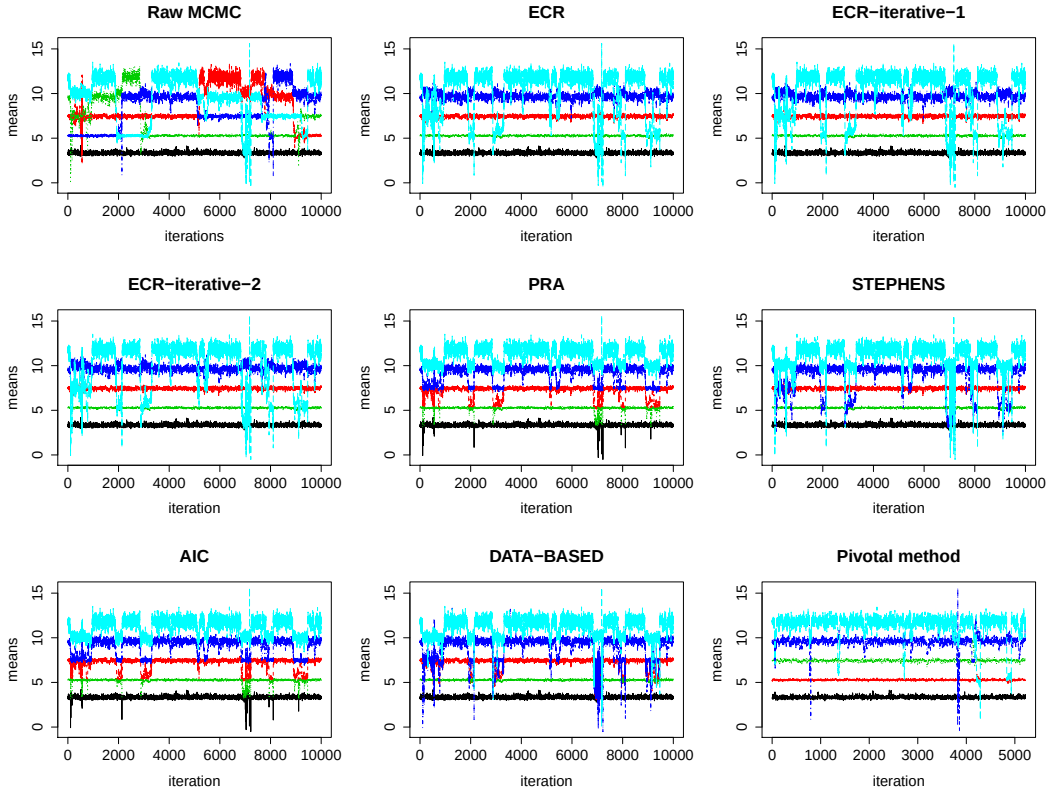


Figure 9: Fishery data. (*From top left to bottom right*) Raw MCMC sample for $\mu_g, g = 1, ..., 5$. Reordered MCMC samples by applying the permutations returned by `label.switching` function, according to methods ECR, ECR-iterative-1, ECR-iterative-2, PRA, STEPHENS, AIC, DATA-BASED. Reordered MCMC samples according to the pivotal method.

In Figure 9 the raw MCMC sample (Top left) and the reordered MCMC samples for $\mu_g, g = 1, ..., 5$, for different methods, are shown. Despite an ordering constraint for the mean components (the priors are chosen according to the `independence` option, which favours a natural ordering of the means), label switching occurs, and the raw sampler is unable to yield useful means estimates for the single components. The `label.switching` function of the same package is used to reorder the obtained chains according to the resulting permutations. The methods from the **label.switching** package seem to perform similarly. In particular, for the greatest mean (light blue trace) there is a global tendency of switching. We note that for the DATA-BASED method the same happens also for the second mean (blue trace). Our pivotal method seems to work better in isolating the five high-posterior density regions. We recall that the reordering for our method is explained by (15).

Concerning the computational times reported in Table 4, AIC is the fastest method, since it only applies an ordering constraint and consequently permutes the simulated MCMC output, while STEPHENS —a probabilistic relabelling— is the slowest. Our method is quite fast, especially if compared with ECR-iterative-1, ECR-iterative-2 and DATA-BASED.

Method	CPU time (sec.)
STEPHENS	344.50
PRA	3.96
ECR	8.66
ECR-iterative-1	60.83
ECR-iterative-2	28.39
AIC	0.08
DATA-BASED	22.68
Pivotal	**9.57**

Table 4: Fishery data: CPU times in seconds for different methods, with $H = 11000$, burn-in=1000, $n = 256$ and $G = 5$.

9 Conclusions

We propose a simple procedure for dealing with label switching in Bayesian mixture models, based on the identification of as many pivots as mixtures components, used for relabelling the resulting MCMC chains. The main novelty of our contribution consists in providing some useful indications on how to choose the pivots, since, as mentioned in Section 6.1, the idea of solving the relabelling issue by fixing the groups for some units is not new (Chung et al., 2004). We suggest to adopt one of six alternative methods based on a maximization or a minimization of some quantities derived from a similarity matrix obtained through the MCMC sample, or a further demanding algorithm suitable when the number of groups G is relatively small (e.g. $G = 4$).

A fundamental issue is represented by the pairwise (perfect) separation between pivots, since it is crucial for the proposed procedure and, usually, non-trivial.

From a computational side, the method appears to be fast and simpler than other relabelling methods, since it does not require a maximization/minimization step at each iteration, and only requires a permutation of the labels induced by the pivots membership. A simulation study is conducted in order to test the proposed solution on different possible scenarios, showing overall good performances. A case study on a real dataset is presented, and the results seem to confirm the advantage of using the proposed methodology. Moreover, when also considering the computational time of our algorithm compared to some procedures available in the **label.switching** `R` package, we conclude that the proposed methodology may represent a valid approach to the label switching problem and, in some cases, may be preferable to other existing solutions.

References

Chung, H., E. Loken, and J. L. Schafer (2004). Difficulties in drawing inferences with finite-mixture models. *The American Statistician 58*(2).

Fritsch, A. and K. Ickstadt (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian analysis 4*(2), 367–391.

Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association 96*(453), 194–209.

Grün, B. (2011). bayesmix: Bayesian mixture models with jags. *R package version 0.7-2, URL http://CRAN. R-project. org/package= bayesmix*.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of classification 2*(1), 193–218.

Jasra, A., C. Holmes, and D. Stephens (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 50–67.

McLachlan, J. and D. Peel (2000). *Finite Mixture Models*. John Wiley & Sons.

Papastamoulis, P. (2016). label.switching: An R package for dealing with the label switching problem in mcmc outputs. *Journal of Statistical Software, Code Snippets 69*(1), 1–24.

Puolamäki, K. and S. Kaski (2009). Bayesian solutions to the label switching problem. In *Advances in Intelligent Data Analysis VIII*, pp. 381–392. Springer.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

Titterington, D. M., A. F. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. Wiley,.

Yao, W. and L. Li (2014). An online bayesian mixture labelling method by minimizing deviance of classification probabilities to reference labels. *Journal of Statistical Computation and Simulation 84*(2), 310–323.

Appendix

MUS algorithm

The algorithm of *Maxima Units Search* is an alternative method for detecting pivots which does not rely upon a maximization/minimization step as the other six procedures in (18) and (19), but it searches for $\hat{G}$ pivots which satisfy a proper feature within the estimated similarity matrix C. The underlying idea is to choose as pivots those units in correspondence of which the $\hat{G}\times\hat{G}$ sub-matrix of C with only the row and columns corresponding to $i_1, \ldots, i_{\hat{G}}$, is more often (close to) the identity matrix. Let us denote this sub-matrix of $C = (c_{ij})$ only containing the rows and columns corresponding to the pivots $i_1, \ldots, i_{\hat{G}}$ by $\mathcal{T}_{(\hat{G}\times\hat{G})}$. It is worth stressing that for a small number of groups (e.g., $G = 4$) and a sample size n ranging between 100 and 1000, this research can be computationally demanding. Furthermore, a positive number of identity matrices is not always guaranteed. However, the MUS algorithm has proved to be efficient in terms of mean square errors for group means estimation, as shown in Table 3. The main steps of the algorithm are summarized below.

(i) For every group g, $g = 1, ..., \hat{G}$, find the *maxima* units $j_g^1, ..., j_g^M$ within matrix C, i.e. the units in group g with the greatest number of zeros in correspondence of the units of the other $\hat{G} - 1$ groups, where M is a precision parameter fixed in advance (in our simulation study $M = 5$).

(ii) For these $M \times \hat{G}$ units, count the number of distinct identity sub-matrices of rank $\hat{G}$ $\mathcal{T}_{(\hat{G}\times\hat{G})}$ which contain them.

(iii) For each group g, $g = 1, ..., \hat{G}$, select the unit which yields the maximum number of identity matrices of rank $\hat{G}$. Such unit represents the pivot to be used for relabelling the chains as explained in Section 6.